Recueil du Symposium 2001 de Statistique Canada
La qualité des données d'un organisme statistique : une perspective méthodologique

COORDINATION D'ÉCHANTILLONS PAR LA MÉTHODE DES MICROSTRATES

Pascal Rivière[1]

RÉSUMÉ

La question de la coordination d'échantillons est primordiale pour les enquêtes-entreprises, car c'est une façon d'étaler le fardeau des enquêtes. Dans bien des méthodes de coordination, les numéros aléatoires qui caractérisent les unités sont permanents et le mode d'échantillonnage varie. Dans la méthode des microstrates, c'est la fonction de sélection qui est permanente. En revanche, les numéros aléatoires font l'objet d'une permutation systématique entre unités à diverses fins de coordination, qu'il s'agisse d'étaler le fardeau des enquêtes, de réduire au minimum le chevauchement entre deux enquêtes ou d'actualiser des échantillons permanents. Les permutations se font aux intersections de strates appelées microstrates. Cette méthode offre de bonnes propriétés mathématiques et une stratégie générale de coordination d'échantillons où les naissances, les décès et les changements de strates font l'objet d'un traitement automatique. Il n'y a aucune contrainte particulière pour les stratifications ni pour les taux de renouvellement des échantillons permanents. On a conçu deux logiciels pour l'application de cette méthode et son évolution future, à savoir SALOMON en 1998 et MICROSTRAT en 2001.

MOTS CLÉS : Stratification; Échantillonnage d'enquête; Coordination d'échantillons; Fardeau; Échantillons permanents; Vecteurs de nombres aléatoires; Permutations

1. INTRODUCTION

La coordination d'échantillons se fixe bien des objectifs dans les enquêtes-entreprises, comme le mentionne Royce (2000). On peut distinguer deux grands buts. Le premier est de gérer le fardeau d'enquête imposé aux entreprises en réduisant au minimum le taux de reprise entre deux échantillons (coordination négative) ou en étalant le fardeau accumulé. Le second est d'actualiser les échantillons permanents, c'est-à-dire de gérer le chevauchement entre deux prises d'échantillon consécutives et les périodes d'inclusion d'unités dans un échantillon permanent, ainsi que les unités en sortie d'échantillon ou en entrée. Cela soulève une foule de questions (coordination entre échantillons d'entreprises et échantillons d'établissements, par exemple), et il est impossible de tenir simultanément compte de tous les types de contraintes.

Les méthodes de coordination d'échantillons ont une longue histoire dans une abondance de pays, qu'il s'agisse de la méthode dite de Jales (Atmer, Thulin et Backlund, 1975), de la méthodologie EDE (De Ree, 1983), du système Océan (Cotton, Hesse, 1992), de l'échantillonnage synchronisé (McKenzie, Gross, 2000) ou de l'échantillonnage par mélange de Poisson (Teikari, 2001). Hesse (1999) expose en détail les principales méthodes employées. La méthode des microstrates que nous présentons ici a été conçue en 1998 dans le cadre du projet SUPCOM de coordination d'échantillons d'Eurostat (Rivière, 1998), le but étant de disposer d'un logiciel général que puisse utiliser le profane pour des coordinations diverses.

Nous expliquerons d'abord ce que nous entendons par coordination d'échantillons et décrirons brièvement les méthodes qui existent, après quoi nous énoncerons le principe d'une sélection séquentielle d'échantillons coordonnés et l'idée d'une permutation de numéros aléatoires en fonction du fardeau

[1] Pascal Rivière, Université de Southampton, Département de statistique sociale, Highfield, Southampton S017 1BJ, ROYAUME-UNI.

d'enquête, ce qui nous amènera à parler de la méthode de permutation dans des « microstrates » selon ce même critère du fardeau d'enquête. Nous décrirons ensuite les grandes caractéristiques du logiciel d'application de cette méthode, en gros du point de vue de l'utilisateur. Nous conclurons par une liste des caractéristiques, des avantages et des inconvénients de cette même méthode.

2. SÉLECTION SÉQUENTIELLE D'ÉCHANTILLONS STRATIFIÉS COORDONNÉS

Dans ce qui suit, il s'agira dans tous les cas d'échantillonnages aléatoires stratifiés simples sans remise (EASSSR), ce qui correspond à la majorité des plans d'échantillonnage effectivement appliqués aux fins des enquêtes-entreprises. D'autres plans de sondage sont possibles et d'autres modes de coordination s'ensuivront. Ainsi, Ohlsson (2000) analyse la coordination des échantillons ppt.

Souvent, il est difficile, dans des enquêtes-entreprises réelles, de disposer d'avance de tous les renseignements voulus sur les futurs plans d'échantillonnage, d'où l'impossibilité d'effectuer une coordination globale de tous les échantillons (tous les échantillons de la même année, par exemple). La coordination n'est alors possible que d'enquête en enquête. Dans une enquête déterminée, notre but sera de prélever un échantillon coordonné avec tous les échantillons antérieurs et de sorte que les probabilités de sélection soient contrôlées. La difficulté est désormais de définir comment on peut constituer des échantillons coordonnés.

Dans une séquence d'échantillons $(s^1,...,s^K)$, tout échantillon s^k sera entièrement caractérisé par un plan d'échantillonnage (qui comporte une stratification), un vecteur de numéros aléatoires et une fonction de sélection.

Plan d'échantillonnage

La stratification se définit comme partition de l'univers : $U = \bigcup_{h=0}^{H_k} U_h^k$, où U_h^k = strate *h* de l'enquête *k*.

Par convention, U_0^k est le sous-ensemble d'unités hors échantillon, dont le taux d'échantillonnage est nul.

La base de sondage de l'enquête *k* est alors $\bigcup_{h=1}^{H_k} U_h^k$. H_k = nombre de strates d'échantillonnage dans l'enquête *k*.

Comme il s'agit d'échantillonnages EASSSR, les plans d'échantillonnage peuvent alors se définir comme la liste des effectifs de toutes les strates. Pour chaque enquête *k* et chaque strate d'échantillonnage *h*, le nombre total d'unités est désigné par N_h^k, et le plan d'échantillonnage indique le nombre d'unités n_h^k à échantillonner dans cette strate.

Nous avons : $|U| = N = \sum_{h=0}^{H_k} N_h^k$, $|s^k| = n^k = \sum_{h=0}^{H_k} n_h^k$, où $n_0^k = 0 \ \ \forall k$

Par définition, les probabilités de sélection sont : $\forall j \in U_h^k, \ \ \pi_j^k = \left(N_h^k\right)^{-1} n_h^k$

Dans un échantillonnage EASSSR, on peut aussi décrire le plan par les taux d'échantillonnage, mais il est alors nécessaire de dégager des nombres d'unités à l'aide de techniques d'arrondis (Cox, 1987). Pour simplifier, nous nous limiterons aux cas où le plan d'échantillonnage se définit par des nombres d'unités par strate.

Le plan k^e d'échantillonnage est dans ce cas caractérisé par : $Q^k = \{(U_0^k,0),(U_1^k,n_1^k),...,(U_{H_k}^k,n_{H_k}^k)\}$

Vecteurs de numéros aléatoires

Dans notre contexte, un « numéro aléatoire » est une valeur de variable aléatoire tirée dans une distribution uniforme [0 , 1) et attribuée à une unité. Chaque unité se voit assigner un numéro aléatoire propre. Un vecteur de numéros aléatoires (VNA) est un vecteur à *N* éléments, qui sont les numéros aléatoires attribués à toutes les unités de *U*. Un tel vecteur est un instrument fondamental de prélèvement d'échantillons. Il est essentiel que ses numéros soient indépendants et identiquement distribués (i.i.d.) si nous entendons échantillonner avec facilité et dans une parfaite maîtrise des probabilités de sélection.

Voilà pourquoi le tout premier VNA ω^1 est tel que tous les numéros aléatoires sont i.i.d. Dans ce cas et contrairement à bien d'autres méthodes, notre méthode comportera des numéros aléatoires qui ne sont pas « permanents ». Avant chaque enquête *k*, on établira un nouveau VNA ω^k par déduction des VNA antérieurs comme nous l'exposerons au prochain chapitre. Pour chaque nouvelle unité (naissance, par exemple), on tire un nouveau numéro aléatoire (et l'attribue à l'unité) indépendamment des numéros aléatoires antérieurs.
Ainsi, avant le prélèvement du k^{e} échantillon, chaque unité *j* a un numéro aléatoire ω_j^k .

Sélection de l'échantillon

L'inclusion d'une unité *j* dans l'échantillon s^k peut être désignée par un indicateur d'appartenance I_j^k , qui prend la valeur 1 si $j \in s^k$ et 0 dans les autres cas. Étant donné le plan d'échantillonnage et le VNA, le prélèvement de l'échantillon *k* par la technique EASSSR exige que l'on élabore une fonction déterministe qui produit le vecteur des indicateurs d'appartenance de sorte que ces indicateurs respectent les contraintes du plan. Il faut aussi garantir son invariance pour toute permutation des éléments à l'intérieur des strates d'échantillonnage. Une façon simple de respecter ces contraintes est de sélectionner dans chaque strate *h* les unités ayant les n_h^k numéros aléatoires les plus bas.

Ainsi, une séquence de prélèvement d'échantillon se définira comme une séquence $\left((Q^1,\omega^1),...,(Q^K,\omega^K)\right)$ telle que chaque VNA ω^k est fonction des VNA antérieurs ou encore un vecteur indépendant de tous les VNA antérieurs et, telle toutes les sélections d'échantillons se font en sélectionnant les unités ayant les numéros aléatoires les plus bas dans chaque strate.

3. PERMUTATIONS EN FONCTION DU FARDEAU D'ENQUÊTE

Remplir un questionnaire d'enquête-entreprises a tout d'une tâche longue et fastidieuse. Nombreuses sont les entreprises qui se plaindront de la quantité de travail qui leur est imposée, ce que l'on appelle la charge ou le fardeau d'enquête. Dans un ensemble quelconque de plans d'échantillonnage, la coordination d'échantillons ne saurait alléger la charge totale (sur l'ensemble des unités et des enquêtes), mais elle peut mieux l'étaler. De Ree (1983) a proposé de trier les unités par fardeau accumulé croissant juste avant le prélèvement de l'échantillon. Ainsi, les unités au fardeau moindre seront plus susceptibles d'être sélectionnées. Nous suggérons de recourir à un tel tri pour une permutation des numéros aléatoires entre unités, ce qui veut dire que ces numéros ne seront pas permanents. Le tri se faisant par fardeau accumulé croissant, il est possible de garantir que moins la charge d'une unité sera lourde, plus ses probabilités de sélection seront grandes.

Fardeau accumulé

Nous définissons un coefficient de fardeau pour chaque enquête. Celui-ci est censé représenter le temps ou le coût qu'exige la réponse au questionnaire. Les coefficients de fardeau peuvent varier selon les enquêtes,

car les coûts en question peuvent se révéler hautement variables. Il peut exister une pluralité de coefficients de fardeau pour une enquête, mais il n'y en aura qu'un pour chaque strate d'échantillonnage. Ces coefficients ne sont pas une estimation réelle du fardeau : une estimation semblable est fort difficile et cette charge comprend un élément subjectif qui ne se prête pas à une mesure. Dans la pratique, les coefficients de fardeau sont des nombres entiers.

Soit c_j^k le coefficient de fardeau de l'unité *j* dans l'enquête *k*. Par définition, le fardeau accumulé de l'enquête *m* à l'enquête *K*, ce que l'on appellera le fardeau accumulé (*m,K*), sera pour l'unité *j* :

$CB_j^{m,K} = \sum_{k=m}^{K} c_j^k I_j^k$ (CB pour « cumulative burden »)

Permutations respectives en fonction du coût et du fardeau

Une permutation des éléments d'un vecteur *X* est dite <u>en fonction du coût</u> s'il existe un vecteur $C = (c_1,...,c_N)$ tel que les éléments du nouveau vecteur *X'* s'ordonnent selon le vecteur coût :

$x_i' < x_j' \Leftrightarrow c_i < c_j \text{ or } (c_i = c_j \text{ and } x_i < x_j)$

Une permutation est dite <u>en fonction du fardeau</u> s'il s'agit d'une permutation en fonction du coût où le vecteur coût est un vecteur de fardeaux accumulés, qui est un vecteur aléatoire. Avec une telle permutation, on s'assure que, dans tout sous-échantillon, les numéros aléatoires sont triés en fonction du fardeau total. Ainsi, les unités à la charge la plus lourde auront les numéros aléatoires les plus élevés dans chaque sous-échantillon.

4. PERMUTATIONS APROPRIÉES DANS DES MICROSTRATES EN FONCTION DU FARDEAU

Principe des microstrates

La grande question est maintenant la suivante : où les tris se font-ils? Si le tri (et la permutation) s'opérait dans les strates d'échantillonnage par exemple, cela soulèverait une question. Supposons, par exemple, que nous prélevons deux échantillons A et B, l'un et l'autre stratifiés par le nombre de salariés. Dans l'échantillon A, toutes les entreprises ayant plus de 20 salariés sont incluses, alors que, dans les entreprises qui ont moins de 20 salariés, la fraction de sondage est de 50 %. Dans l'échantillon B, nous avons une seule strate formée d'entreprises comptant de 10 à 49 salariés. Supposons maintenant qu'une permutation par fardeau accumulé croissant se fait dans la strate d'échantillonnage B. Il est évident que les unités ayant de 20 à 49 salariés se situeront à la fin de [0, 1] et que nous aurons alors dans le prochain échantillon une surreprésentation des entreprises ayant de 10 à 19 salariés. En d'autres termes, les probabilités de sélection seront moindres dans la tranche 20-49. Par intuition, on peut voir qu'une permutation dans la tranche 10-49 ne convient pas, car les unités des tranches 10-19 et 20-49 n'ont pas les mêmes « antécédents » sur le plan des strates d'échantillonnage.

La première façon de résoudre le problème est d'obliger l'utilisateur à créer des strates d'échantillonnage à l'aide d'une partition élémentaire, c'est-à-dire de ce que Van Huis et coll. (1994) appellent des unités de base. L'utilisateur de ce système de coordination n'est pas libre de son choix de strates d'échantillonnage, celles-ci devant être des combinaisons d'unités de base. La technique est efficace, mais pose un problème si le but est d'élaborer un système général de coordination où ce même utilisateur devrait pouvoir définir la stratification qu'il recherche sans contraintes préalables. Ajoutons que, même si les strates d'échantillonnage se définissent par combinaison d'unités de base, il ne suffit pas de tenir compte des naissances ni des changements de strates d'une manière générale.

L'idée à l'origine des microstrates est de laisser l'utilisateur libre de choisir sa stratification et d'avoir quelque chose d'assez général pour les naissances et les changements de strates. Le principe d'ensemble est fort simple : au lieu d'imposer au départ des unités de base qui sont des sortes d'intersections de strates d'échantillonnage, nous essaierons de déduire notre stratification de la stratification des échantillons antérieurs, quelle qu'en soit la nature. Les microstrates sont simplement des intersections de strates antérieures d'échantillonnage. Si nous voulons étaler le fardeau accumulé pour les enquêtes *m* à *K*, les microstrates seront définies comme l'intersection des strates d'échantillonnage de toutes les enquêtes en question. Chaque microstrate *A* sera alors :

$A = U_{h_{m1}}^{m} \cap ... \cap U_{h_K}^{K}$, où $0 \le h_m \le H_m, ..., 0 \le h_K \le H_K$

Séquences appropriées d'échantillons coordonnés en fonction du fardeau

Avec une séquence d'échantillons, une permutation (*m*,*K*) est simplement une permutation de ω^K dans les microstrates (*m*,*K*) où des tris se font par fardeau accumulé croissant (*m*,*K*). Une telle permutation sera dite appropriée au sens que les échantillons servant au calcul du fardeau accumulé seront les mêmes que les échantillons servant à la microstratification, c'est-à-dire que le fardeau accumulé et la microstratification auront le même échantillon initial *m* et le même échantillon final *K*.

Les permutations appropriées sont à la base même de la méthode des microstrates. L'idée est de procéder à des permutations appropriées en fonction du fardeau après chaque prélèvement d'échantillon, ce qui permet d'étaler le fardeau accumulé. Pour chaque permutation, il serait possible de réattribuer au besoin tous les coefficients de fardeau de toutes les unités. La seule contrainte serait que, pour une enquête *k* déterminée, ces coefficients doivent être constants sur les strates d'échantillonnage de *k*.

Une séquence d'échantillons coordonnés en fonction du fardeau est donc une séquence d'échantillons $\left((Q^1, \omega^1), ..., (Q^K, \omega^K)\right)$ telle que :

- ω^1 est un vecteur aléatoire i.i.d.;
- $\forall k, 1 < k \le K, \exists m_k, q_k, 1 \le m_k \le q_k < k$ / ω^k est une permutation (m_k, q_k) de ω^{q_k}

Principale propriété des séquences d'échantillons coordonnés en fonction du fardeau

Rivière (2001) démontre que, dans une séquence d'échantillons en fonction du fardeau où l'échantillon de début est le même dans toutes les permutations selon le critère du fardeau (ce qui signifie que $m_k = m \ \forall k$), tous les vecteurs aléatoires ω^k sont i.i.d.

Ainsi, si nous recourons toujours à des permutations appropriées à partir d'une « première » enquête donnée, nous sommes sûrs que les probabilités de sélection de tout ordre sont intégralement prises en compte.

Il convient de noter que, dans une telle orientation d'étalement de fardeau, la coordination se fait d'échantillon en échantillon (chacun étant coordonné avec les échantillons antérieurs en vue de réduire le fardeau total au minimum), et non pas sur plusieurs échantillons simultanément. C'est là un choix délibéré, puisque, dans une foule d'instituts statistiques nationaux, on ne connaît pas longtemps d'avance les plans d'échantillonnage des enquêtes-entreprises.

Indépendance entre les microstrates et le prélèvement d'échantillons

Une des grandes caractéristiques est que les tris ne se font pas à l'étape du prélèvement de l'échantillon. Préalablement à l'échantillonnage et avant même de connaître le plan d'échantillonnage pour le prochain échantillon à constituer, on permute les numéros aléatoires dans chaque microstrate pour être sûr que plus le numéro aléatoire d'une unité sera élevé, plus grand sera aussi son fardeau accumulé. Les microstrates sont alors les intersections des strates d'échantillonnage des enquêtes antérieures, mais elles ne tiennent manifestement pas compte des strates d'échantillonnage pour l'échantillon à prélever.

Ainsi, le prélèvement de l'échantillon même n'a absolument rien à voir avec les microstrates. Étant donné le vecteur de numéros aléatoires et les plans d'échantillonnage, l'exercice est totalement déterministe, consistant à sélectionner les numéros aléatoires les plus bas dans chaque strate. Cela veut dire que la fonction de sélection est toujours la même, comme nous l'avons déjà précisé. On s'écarte donc pour l'essentiel des méthodes fondées sur des numéros aléatoires permanents (comme la méthode dite de JALES) où ces numéros sont invariables et la fonction de sélection, variable (par exemple, lorsque la sélection est définie par un point de départ dans [0, 1) et une direction). Dans la méthode des microstrates, les numéros aléatoires changent toujours, mais la fonction de sélection est fixe.

5. EMPLOI DE LA MÉTHODE DANS LA PRATIQUE : PLURALITÉ DE PERMUTATIONS

La méthode fondée sur les microstratifications peut présenter un grand inconvénient, car plus le fardeau s'accumule dans le temps, plus les microstrates diminuent. Dans le cas de microstrates minuscules, les permutations ne sont pas très utiles, et il est clair qu'elles perdent tout leur sens dans des microstrates formées d'une seule unité. Voilà pourquoi on se doit d'améliorer la technique de permutation pour pouvoir l'appliquer dans la pratique. Comment peut-on concevoir une méthode qui, faisant appel au principe des microstrates, n'en évite pas moins la question de la taille de ces microstrates? En réalité, l'idée est de procéder à plusieurs tris à plusieurs niveaux. Dans ce qui suit, nous distinguerons une coordination négative globale (visant à étaler le fardeau accumulé) d'une coordination spécifique avec un autre échantillon (qu'elle soit positive ou négative).

Coordination négative globale

Dans la technique du fardeau accumulé que nous avons mentionnée, l'échantillon de départ de l'accumulation joue un grand rôle. Pour éviter l'obstacle de la petite taille des microstrates, nous effectuerons plusieurs permutations appropriées en fonction du fardeau avec différents échantillons de départ et donc différentes tailles de microstratifications.

Supposons que nous avons prélevé K échantillons. Le problème est alors d'élaborer un VNA ω^{K+1} qui servira au prélèvement du K+1^{e} échantillon. Pour ce faire, on doit prendre le dernier VNA ω^{K}, puis faire trois permutations (m,K) avec différentes valeurs de m :

- la première a lieu dans les strates du K^{e} échantillon avec tri selon l'ordre croissant de l'indicateur d'appartenance à l'enquête K;
- la deuxième a lieu dans les microstrates, à partir de la première enquête de l'année en cours y, avec tri selon le fardeau accumulé croissant depuis le début de l'année;
- la troisième a lieu dans les microstrates, à partir de la première enquête de l'année y-2, avec tri selon le fardeau accumulé croissant depuis le début de l'année y-2.

Nous pouvons voir l'avantage de cette méthode, le premier tri devant permettre un renouvellement régulier des unités dans de grandes strates (les plus grandes possible en fait, puisqu'il s'agit des strates d'échantillonnage). Toutefois, ce premier renouvellement ne tient pas compte des fardeaux accumulés et rappelle l'esprit de la technique de renumérotation du système Océan (Cotton, Hesse, 1992). Le deuxième tri permet un renouvellement en fonction des fardeaux accumulés dans des microstrates qui ne sont pas

encore trop petites. La troisième permutation est la plus importante pour l'étalement du fardeau d'enquête. Elle permet de gérer le fardeau accumulé sur une assez longue période en gardant à l'esprit que, ainsi que nous l'avons souligné, les permutations n'ont plus guère d'effet s'il s'agit de microstrates minuscules.

Ce tri produit un nouveau VNA ω^{K+1}. À ce stade, nous ne savons pas nécessairement ce que sera le plan d'échantillonnage de l'enquête *K+1*. Lorsque nous le connaîtrons, le prélèvement de l'échantillon se fera de la même façon, c'est-à-dire que, dans chaque strate *h*, nous sélectionnerons les unités n_h ayant les numéros aléatoires les plus bas (n_h donnés par le plan d'échantillonnage).

Exemple de microstrates

Considérons une population de 33 unités. Supposons qu'il y a trois enquêtes, les deux premières ayant deux strates d'échantillonnage 1 et 2 (plus les unités hors échantillon) et la troisième en a 3 (plus les unités hors échantillon). Chaque enquête définit une partition de la population et, pour chaque enquête et chaque unité, le numéro de strate est donné.

Tableau 1 : Décomposition de la population en microstrates

Enquête 1	00	1	11 22	11 222	0	0 11111 2222	1 2	111 22222
Enquête 2	00	0	11 11	22 222	0	1 11111 1111	1 1	222 22222
Enquête 3	00	1	11 11	11 111	2	2 22222 2222	3 3	333 33333
Microstrates	**1**	**2**	**3 4**	**5 6**	**7**	**8 9 10**	**11 12**	**13 14**

Il y a 14 microstrates (1,3), c'est-à-dire d'intersections des strates des trois enquêtes (1 à 14). Il y a 8 microstrates (2,3), c'est-à-dire d'intersections des deuxième et troisième enquêtes (colonnes du tableau 1 : {1}, {2}, {3,4}, {5,6}, {7}, {8,9,10}, {11,12}, {13,14}). Il y a enfin 4 microstrates (3,3), qui sont les strates d'échantillonnage de la troisième enquête : {1}, {2-6}, {7-10} et {11-14}.

Coordination spécifique

Dans notre méthode, chaque prélèvement d'échantillon doit recevoir l'apport d'une stratification, d'un plan d'échantillonnage et de caractéristiques de coordination. Une coordination négative globale est une façon possible de coordonner des échantillons (c'est donc une caractéristique possible de coordination), mais il y en a d'autres.

L'utilisateur d'un système de coordination d'échantillons pourrait aussi vouloir coordonner son enquête avec une autre en mode positif ou négatif. Cette coordination peut encore mener à l'exécution de tris dans des partitions. Au terme de ces tris quelconques, un numéro aléatoire sera toujours assigné à chaque unité. Par l'emploi de ce numéro aléatoire, l'échantillon est ensuite toujours prélevé de la même manière. Décrire la coordination, c'est donc décrire les permutations qui précèdent.

Dans la méthode retenue, le point de départ est l'aveu de l'impossibilité d'atteindre tous les objectifs simultanément : pour un échantillon donné, nous ne pouvons obtenir à la fois une bonne coordination spécifique et un échantillon où il y a étalement du fardeau. Le prélèvement de l'échantillon étant toujours le même (étant donné le VNA), le problème est de trouver un VNA ω^{K+1} i.i.d. qui nous garantira la coordination spécifique que nous recherchons.

Soucieux de ne pas mêler coordination globale et coordination spécifique, nous n'aurons pas à exécuter les trois premières permutations (voir 5.1). L'idée est alors simple : pour coordonner le K+1^{e} échantillon A avec l'échantillon B, nous prendrons d'abord le VNA ω^{K_B} *utilisé pour le prélèvement de l'échantillon B.* Dans ce cas :

- si nous désirons une coordination positive des échantillons A et B (100 % de reprise), $\omega_j^{K+1} = \omega_j^{K+1}\ \forall j$
- si nous désirons une coordination négative de A et B, $\omega_j^{K+1} = 1 - \omega_j^{K+1}\ \forall j$.

Cela veut dire qu'il n'y a pas de permutations du tout. Il est facile de constater que ce mode de coordination spécifique équivaut strictement à la technique utilisée dans la méthode de JALES où nous sommes parfaitement maîtres du taux de reprise si la définition de la stratification ne change pas et que les unités demeurent dans leurs strates initiales.

Coordination positive avec taux de renouvellement et actualisation de l'échantillon permanent

Dans une telle coordination, le but est de veiller à ce que, d'un échantillon à l'autre, une proportion *r* d'unités soient déséchantillonnées et une proportion 1-*r* demeurent en échantillon. Il s'agit dans le premier cas de ce que l'on appelle le taux de renouvellement et, dans le second, le taux de reprise. Notre façon de procéder peut être la même que dans ce qui précède où *r* était 0 ou 1.

Nous prenons donc d'abord les numéros aléatoires utilisés pour le prélèvement de l'échantillon B. Abstraction faite des naissances, les unités appartenant à l'échantillon B ont les numéros aléatoires les plus bas dans les strates d'échantillonnage B. Pour une coordination avec un taux de reprise *r*, la technique ressemblerait (sans être identique) à la précédente, sauf que la permutation s'opérerait dans les strates d'échantillonnage B, et non pas dans les microstrates.

Considérons une strate d'échantillonnage h comptant N_h unités dont n_h ont été sélectionnés dans l'échantillon B. L'idée est alors de prendre les $r.n_h$ premières unités et de les mettre à la fin de l'intervalle [0, 1), ce qui revient à faire glisser vers le bas les numéros aléatoires des $N_h - r.n_h$ unités restantes. À cause des naissances et des décès, il n'est pas sûr que les premières unités n_h étaient toutes dans l'échantillon B. Par ailleurs, n_h est fixe (voir le plan d'échantillonnage B).

Exemple : 1 strate, 12 unités, taux d'échantillonnage de 50 % pour B, taux de renouvellement 1/3.

Avant la permutation :

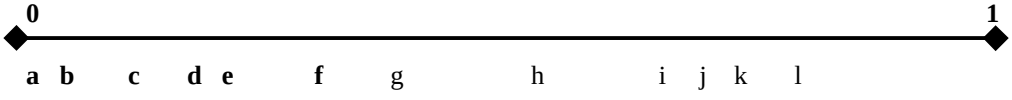

Figure 1 : Positions des numéros aléatoires initiaux des unités

On peut voir que, dans l'enquête B, les unités a, b, c, d, e et f se trouvaient dans l'échantillon. Après permutation, nous obtenons :

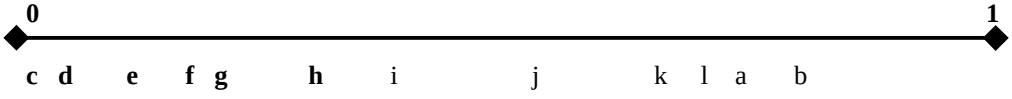

Figure 2 : Positions après permutation

Les unités qui demeurent dans l'échantillon sont maintenant c, d, e et f, les nouvelles unités sont toujours g et h, et a et b quittent l'échantillon. Si on voit [0, 1) comme un cercle, et non pas comme une bande, il est clair que la technique consiste à faire glisser les unités vers le bas le long du cercle, ce qui est parfaitement adapté aux échantillons permanents. Dans la pratique, la nécessité d'un calcul d'arrondis rend la chose un peu plus complexe.

Échantillons permanents

Un échantillon permanent est en fait une suite d'échantillons, chacun étant coordonné avec le précédent avec un taux de reprise donné. Dans le traitement d'échantillons permanents, il importe aussi de gérer les périodes d'inclusion d'unités.

On peut voir dans les échantillons permanents un cas d'espèce de la coordination positive à taux de reprise donné, aussi ceux-ci n'ont-ils rien de particulier dans la méthode que nous proposons. Cela implique que la stratification peut varier d'échantillon en échantillon et que le taux de reprise n'est pas fixe. Par ailleurs, on ne fait rien de précis pour gérer les périodes d'inclusion. Il s'agit d'un contrôle implicite. Ainsi, si le taux d'échantillonnage est fixe, que le taux de renouvellement est $1/p$ (p étant un nombre entier), qu'il n'y a ni naissances, ni décès, ni changements de strates et que si nous oublions la question des arrondis, il est possible de démontrer que la période maximale d'inclusion est p.

Traitement des naissances, des décès et des changements de strates

Dans bien des méthodes de coordination, les naissances, les décès et les changements de strates sont très difficiles à traiter. Dans la méthode des microstrates, la généralité du principe garantit que la question des naissances sera automatiquement traitée, c'est-à-dire sans qu'on ait à intervenir : par convention, une unité nouvelle appartenait dans les enquêtes antérieures à une strate bien particulière, la strate « hors échantillon » (où le taux d'échantillonnage est nul).

Supposons donc qu'une microstrate contient des unités nouvelles. Comme elle se définit comme l'intersection de strates d'échantillonnage antérieures, ces unités appartenaient auparavant toutes à la strate « hors échantillon » et, par conséquent, elles sont toutes des « naissances ». Le principe des microstrates mène à des créations automatiques de « microstrates de naissances ». L'utilisateur n'a pas à s'en soucier, puisqu'une microstrate de naissances est implicitement constructe. La même constatation vaut pour les changements de strates : les microstrates correspondantes se créent automatiquement. Enfin, le traitement des décès ne pose aucun problème particulier.

Bien sûr, naissances, décès et changements de strates ont bien des effets négatifs sur le plan de la gestion de l'étalement du fardeau et de celle du chevauchement (échantillons permanents ou coordination négative). Mais ce que nous voulons faire valoir ici, c'est qu'on n'a pas à ajouter à la méthode, car il n'y a rien de nouveau à concevoir pour le traitement de ces cas. Ajoutons que celui qui emploie la méthode n'a pas à s'en soucier.

6. APPLICATION DE LA MÉTHODE

En 1998, on a élaboré le logiciel SALOMON dans le cadre du projet SUPCOM d'Eurostat par la méthode des microstrates (Rivière, 1998, 1999). Le programme présente plusieurs caractéristiques : définition de la structure de la base de sondage, actualisation de cette base, définition du champ d'observation et de la stratification d'une enquête, choix du type de coordination (négative globale, négative, positive, positive avec taux de reprise, nulle), prélèvement de l'échantillon. Ce logiciel disponible en micro-informatique a été donné par Eurostat en 1999 à un certain nombre d'instituts statistiques nationaux qui l'ont utilisé,

notamment en Europe de l'Est. Ce logiciel conçu en moins d'un an présente certains inconvénients. Il peut être d'un fonctionnement lent à cause de la masse de permutations à opérer.

Au lieu d'améliorer ce logiciel, l'INSEE a décidé d'en élaborer un tout nouveau, MICROSTRAT, qui n'a rien de commun avec SALOMON pour ce qui est des programmes informatiques. L'idée était de résoudre bien des questions comme celle de la coordination spécifique (dans SALOMON, on utilisait une autre méthode de coordination spécifique qui s'est révélée moins efficiente) et de refaire entièrement le mode de stockage de données en vue d'optimiser les permutations. Nous décrirons les grandes caractéristiques communes à ces deux instruments.

Dans le système, les variables de stratification sont modifiées au moins une fois par an : il y a une « actualisation annuelle principale » qui se fait généralement au début de janvier ou à la fin de décembre, après quoi les microstratifications à la base des deux premiers tris (de la coordination négative globale) sont automatiquement renouvelées. C'est que le « fardeau accumulé depuis le début de l'année *y* » (ou *y*-2) n'est pas le même et ne correspond pas aux mêmes stratifications, car l'« année *y* » (ou *y*-2) a connu des changements. On peut procéder à d'autres actualisations dans l'année, mais elles n'influent pas sur la définition des microstrates.

Le logiciel offre une interface pour la définition interactive des strates et du plan d'échantillonnage. Après délimitation du champ d'observation et de la stratification, des codes d'identification de strates (qui sont dans la pratique des nombres entiers) sont automatiquement calculés, puis attribués à toutes les unités. Ces numéros de strate seront des entiers positifs successifs et le numéro 0 sera assigné par convention à la strate « hors échantillon ». La délimitation du champ d'observation par l'utilisateur devrait implicitement mener à l'attribution du numéro 0. Le calcul du numéro de strate de chaque unité se fait pour chacun des échantillons. Le logiciel a alors dans sa base de données les numéros de strate de toutes les unités et de toutes les enquêtes depuis le début de l'année *y*-2. Lorsqu'une nouvelle unité est introduite dans la population (naissance), les numéros de strate de l'unité dans les enquêtes antérieures sont automatiquement fixés à 0 (c'est une attribution automatique qui est invisible à l'utilisateur).

Le mode de coordination par défaut est alors la « coordination globale », sinon l'utilisateur précise avec quelle enquête il désire une coordination positive (ou négative). Sauf pour la définition des coordinations spécifiques (dans le cas ou non d'échantillons permanents), il n'a à se soucier de rien en matière de coordination et n'a pas même à connaître l'existence de microstrates. La construction de microstrates et les tris effectués font partie de la mécanique interne du logiciel, de son moteur, et n'offre aucun intérêt particulier pour l'utilisateur. Toutes les indications nécessaires (numéro de strate, présence dans l'échantillon ou non et numéro aléatoire pour chaque unité et chaque enquête) sont stockées pour trois ans.

7. CONCLUSION

Le système proposé est général. Il repose sur quatre principes :

1) pour faire la coordination d'échantillons, nous pouvons, au lieu d'agir sur la fonction de sélection, modifier les vecteurs de numéros aléatoires et ne rien changer à la fonction de sélection;
2) pour l'étalement du fardeau total, l'idée est de permuter les éléments des vecteurs de numéros aléatoires par tri selon les fardeaux accumulés dans des sous-populations appropriées que l'on appelle microstrates;
3) la coordination spécifique se fait par réutilisation des numéros aléatoires de l'enquête avec laquelle il y a coordination; pour une coordination spécifique à taux de reprise r ($0<r<1$), il faut une permutation déterministe supplémentaire;
4) les échantillons permanents sont un cas d'espèce de la coordination positive spécifique à taux de reprise; les périodes d'inclusion font l'objet d'une gestion implicite par opposition à un traitement explicite.

Ainsi, du point de vue de l'utilisateur, la première grande caractéristique d'un instrument d'application de cette méthode (SALOMON ou MICROSTRAT) est que l'intéressé est libre de choisir les stratifications, les taux de reprise (dans le cas des échantillons permanents) et les types de coordination. La seconde est que la méthode n'exige aucun traitement particulier des naissances, des changements de strates ni des unités qui entrent en échantillon ou en sortent : de nombreuses microstrates sont créées, mais l'utilisateur n'a pas à s'en soucier. Il importe néanmoins de signaler que les besoins de coordination sont nombreux (étalement de fardeau, échantillons permanents, coordination spécifique) et que, quel que soit l'instrument dont on dispose, il est impossible de répondre à tous simultanément : le méthodologiste doit juger des besoins les plus importants.

REMERCIEMENTS

J'aimerais remercier Chris Skinner, Robert Clark et Patrick Hernandez des nombreuses discussions fécondes que nous avons eues.

BIBLIOGRAPHIE

Atmer, J.G., Thulin G., et Backlund S. (1975). Coordination of samples with the JALES technique, *Statistik Tidskrift*, 13, pp. 443-450.

Cotton F., Hesse C. (1992). Tirage coordonné d'échantillons stratifiés. *Recueil du Symposium 1992 de Statistique Canada*

Cox L. H. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82, 520-524.

De Ree J. (1983). A system of co-ordinated sampling to spread response burden of enterprises. Présenté par le Netherlands Central Bureau of Statistics (CES/AC.48/43) à la Conference of European Statisticians (Organisation des Nations Unies).

Hesse C. (1999) : Sampling co-ordination : a review by country, *INSEE, Working paper E9908*

McKenzie R., Gross B. (2000). Synchronised sampling. *Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association.

Ohlsson E. (2000). Co-ordination of pps samples over time, *Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association.

Rivière, P. (1998) : Description of the chosen method, *deliverable 2 of SUPCOM 1996 project (part « co-ordination of samples »), pp11-33, Eurostat*

Rivière, P. (1999): Co-ordination of samples: the microstrata methodology, *Proceedings of the 13th Roundtable on Business Survey Frames, INSEE, Paris*

Rivière, P. (2001): Random permutations of random vectors as a way of co-ordinating samples, Working paper of the university of Southampton, June 2001

Royce D. (2000). Issues in coordinated sampling at Statistics Canada, *Proceedings of the Second International Conference on Establishment Surveys, American Statistical Association*.

Teikari I. (2001) Poisson mixture sampling in controlling the distribution of response burden in longitudinal and cross-section business surveys. *Tutkimuksia Forskningsrapporter Research Reports 232, Statistics Finland*

Van Huis M., Koeijers E. and De Ree J. (1994). Response burden and co-ordinated sampling for economic surveys. *Netherlands Official Statistics*, volume 9.